Birlesik Dünya Arastırma BD CENTER Innovasyon ve Yayıncılık Merkezi

Cypriot Journal of Educational Sciences

Volume 17, Issue 4, (2022) 1258-1268

CJES Cypriot Journal of Educational Sciences

www.cjes.eu

Do scoring techniques and number of choices affect the reliability of multiple-choice tests in elementary schools?

Herwin Herwin [a *], Universitas Negeri Yogyakarta, Faculty of Education, Yogyakarta, 55281, Indonesia

https://orcid.org/0000-0002-8882-5087

Triyanto Pristiwaluyo [b], Universitas Negeri Makassar, Faculty of Education, Makassar, 90222, Indonesia

https://orcid.org/0000-0001-8902-5877

Ruslan Ruslan [c], Universitas Negeri Makassar, Faculty of Mathematics and Natural Sciences, Makassar, 90222, Indonesia https://orcid.org/0000-0002-7922-7819

Shakila Che Dahalan [d], Universiti Pendidikan Sultan Idris, Faculty of Human Science, Perak, 35950, Malaysia

https://orcid.org/0000-0001-9841-9392

Suggested Citation:

Herwin, H., Pristiwaluyo, T., Ruslan, R., & Dahalan, S. C. (2022). Do scoring techniques and number of choices affect the reliability of multiple-choice tests in elementary schools?. *Cypriot Journal of Educational Science*. *17*(4), 1258-1268 https://doi.org/10.18844/cjes.v17i4.7149

Received from January 10, 2022; revised from March 22, 2022; accepted from April 25, 2022.

Abstract

The application of multiple-choice tests often does not consider the scoring technique and the number of choices. The study aims at describing the effect of the scoring technique and numerous options towards the reliability of multiple-choice objective tests on social subjects in elementary school. The study is quantitative research with experiments using factorial design. The population of the study was all the items of the standard of multiple-choice objective test through the test development procedure. The sample of the study was part of the items taken randomly through repeated simulations. Data collection was participants' scores taken from the test and repeated simulations 120 times for the analysis. Data were analysed in descriptive quantitative and inferential through the analysis of variance. The results of the study reveal that there is an interaction between scoring technique and numerous options towards the reliability of multiple-choice objective tests on social subjects in elementary school.

Keywords: Scoring technique, numerous options, test reliability, multiple-choice, elementary schools.

* ADDRESS FOR CORRESPONDENCE: Herwin Herwin, Universitas Negeri Yogyakarta, Faculty of Education, Jalan Colombo No. 1 Karangmalang, 55281, Yogyakarta, Indonesia

E-mail address: herwin89@uny.ac.id / Tel.: +6282347789847

1. Introduction

The phenomenon of test implementation all this time, notably multiple-choice objective forms, had been giving us the notion that evaluators are only willing to focus on the role of an answer or response of the test participants, if answered correctly. The role of the questions that are answered incorrectly or left blank by the respondents did not seem be considered even it is ignored. All this time, scoring techniques only depend on the answers of respondents who answered correctly, while questions answered incorrectly, or questions blanked seem to be ignored and both are not scored or the score is 0.

The illustration above influences evaluation activities in the field, so there was variety of cases found in the field. It described that teacher always give directions to students to not blank the answer sheet, teachers tend to teach students so as not to leave any of the questions, all the questions should be filled with answers, although there are some items that may not be understood at all or are not mastered by the students who are participants of the test.

Certainly, this phenomenon is still far from what is expected, if we want to think about the actual measurement of the concept of implementation. Hence, there must be studies to find out a review of theoretical and practical attempts to control or restrain events as noted previously. In this case, the researchers tried to highlight the scoring technique. It is alleged that the scoring technique has an important role in controlling the attitude of respondents in a test. The latitude and flexibility rules of scoring on objective tests can allow respondents to act without obstruction in expressing their answer in a test.

One of the scoring techniques in educational measurement is the concept of scoring without correction technique. Scoring without correction technique is a famous technique and it is most likely used in the practice of scoring multiple-choice test instruments over the years. The scoring without correction technique is relatively simple when compared to other techniques that only scored by summing respondents' correct answers (Bereby-Meyer et al., 2002; Kurz, 1999; Lesage et al., 2013 Mardapi, 2012). Furthermore, Susetyo (2011) also said that scoring without correction technique is applied to give a score of 1 for test items which are answered correctly and a score of 0 for items which are answered incorrectly.

In addition, in the scoring without correction techniques that have been proposed there is a technique also known as the scoring with a penalty. According to Karandikar (2010), the penalty scoring technique or the scoring with correction is a technique which is applied based on his thinking which is to neutralise the possibility of obtaining the correct answer because of presumption. The purpose of the test is to determine the status or potential ability and knowledge of students, who do not have the ability to guess the correct answer. The scoring penalty technique is applied based on the respondents' guesses, which should be given a penalty by reducing the respondent's scores and the indicator of guessing the incorrect answer should be given the punishment or penalty. The scoring penalty technique has its own characteristics, such that there is punishment or a penalty that is imposed on participants for their incorrect answers. Therefore, in the scoring penalty technique, there is not only the role of the respondents' correct answer but also the incorrect answer has role to be accounted in deciding the determination of the final score on the test results of each student who is a participant in the test.

In the multiple-choice objective test, the implementation of scoring without the correction technique and scoring penalty technique is not separate with the role of the answer options that are characteristic of the test form. All respondents' flexibility in guessing the answers on scoring without correction

techniques and all respondents' risks in guessing the answers on the scoring penalty technique tend to depend on the number of answer options that is available in each stem or main part of the question. Therefore, it is expected that evaluators should consider many options that will be used in each item in applying both techniques.

Multiple-choice tests have special characteristics, namely the existence of choices, and participants must choose one of the correct choices (Butler, 2018; Rintayati et al., 2021). Multiple-choice tests are one of the most widely used assessment methods. When faced with questions in multiple-choice tests, a test developer must evaluate each use of the option and choose the most appropriate one (Annie & Chan, 2009). The use of unfavourable options will have an impact on the quality of the test.

One of the important aspects of the test kit is the quality related to reliability. Reliability is an index of stability and consistency in a measuring instrument (Herwin & Nurhayati, 2021; Nájera Catalán & Gordon, 2020). A good measuring instrument is an instrument that tends to give the same (consistent) results even though it is used at different times (Herwin & Mardapi, 2017; Herwin et al., 2022; Otaya et al., 2020; Ziegler & Detje, 2013). This is the basis for the importance of paying attention to the reliability of a test.

Based on the description, the researcher will examine further the level of test stability that is related to the implementation of scoring without correction techniques and scoring penalty technique by controlling the answer option on an objective test instrument, especially multiple-choice objective test in social subject to students. Hence, based on the background that has been stated previously, the researcher is encouraged to conduct the research and conduct experimental studies to compare the test stability of the implementation of scoring without correction techniques and scoring penalty technique by controlling the answer option on a multiple-choice objective test in social subject. The sample of the research comprised of sixth-grade students in elementary school. Based on the background stated, the following research question was asked: Is there an interaction between scoring techniques and the number of options for multiple-choice objective test reliability in elementary school social studies subjects?

2. Method

2.1. Types of research

This study uses a quantitative approach. This type of research is an experimental study with a 2 × 2 factorial design. In this design, there are two independent variables, namely the scoring technique and the number of answer choices. The dependent variable is the reliability of the multiple-choice test. Table 1 presents the research design carried out in this study.

Table 1. Research design

Scoring technique (*A*) / Numerous options (*B*)	Penalty (A_1)	Without correction (A_2)
3 Options (B_1)	A_1B_1 (μ_1)	A_2B_1 (μ_2)
4 Options (B_2)	A_1B_2 (μ_3)	A_2B_2 (μ_4)

2.2. Setting and research subject

This research was conducted on sixth-grade students in elementary school. This study focused on the social subject. The population of the research was all items of multiple-choice objective tests that have been standard through the development of test procedures. The sample of some of the items was taken randomly through repeated simulations. The variables of this research are scoring techniques and numerous options as the independent variable and the reliability of multiple-choice objective tests as the dependent variable.

2.3. Data collection technique

The instrument is a multiple-choice objective test consisting of two forms, i.e., an instrument with three answer options and an instrument with four answer options. The data of this research were collected by the test and data analysis was conducted by repeated simulations for 120 times. For the purposes of validity and reliability, the multiple-choice test instrument used was first submitted to the expert to assess the validity of its content. The results of this assessment indicated that the instrument had met content validity. In addition, the instrument was tested in the field to obtain reliability information.

2.4. Data analysis technique

The data analysis technique used in this research was a descriptive quantitative analysis that is central tendency and inferential analysis of variance. The formulation of the statistical hypotheses in this study is as follows:

(1) $H_0 : \mu A_1 = \mu A_2$

$H_1 : \mu A1 \neq \mu A2$

(2) $H_0 : \mu B1 = \mu B2$

$H_1 : \mu B1 \neq \mu B2$

(3) $H_0 : A \times B = 0$

$H_1 : A \times B \neq 0$

(4) $H_0 : \mu A1B1 = \mu A2B1$

$H_1 : \mu A1B1 \neq \mu A2B1$

(5) $H_0 : \mu A1B2 = \mu A2B2$

$H_1 : \mu A1B2 \neq \mu A2B2$

(6) $H_0 : \mu A1B1 = \mu A1B2$

$H_1 : \mu A1B1 \neq \mu A1B2$

(7) $H_0 : \mu A2B1 = \mu A2B2$

$H_1 : \mu A_2B_1 \neq \mu A_2B_2$

3. Results and discussion

The results showed that the first hypothesis testing states that, ‘there is a difference between the reliability of multiple-choice tests is scored by using scoring penalty technique and multiple-choice tests

are scored by using scoring without correction techniques'. The result of the calculation is that *p*-value = 0.000, at a significance level of 0.05 with H_0 rejection criteria where *p*-value is lesser than *α*-value. Therefore, *p*-value = 0.000 is lesser than *α* = 0.05, and it was decided that the research hypothesis was tested. This means that there are significant differences in reliability between multiple-choice tests scored by using the scoring penalty technique and multiple-choice tests scored by using scoring without correction techniques.

The second hypothesis testing states that, 'there is difference between the reliability of multiple-choice test that uses three answers options and multiple-choice test that uses four answers options'. The result of the calculation is that *p*-value = 0.000, at a significance level of 0.05 with H_0 rejection criteria where p-value is lesser than α-value. Therefore, *p*-value = 0.000 is lesser than *α* = 0.05, and it was decided that the research hypothesis was tested. This means that there are significant differences in reliability between the multiple-choice test that uses four answer options and the multiple-choice test that uses four answer options.

The third hypothesis testing states that, 'there is interaction between numerous options and scoring techniques towards the reliability of multiple-choice objective test'. The test used is *F*-test. The analysis data obtained the coefficient of F_0 is 1.781, degrees of freedom (df) is 3/166 and *p*-value is 0.000. Because the coefficient *p*-value is lesser than *α* = 0.05, the research hypothesis is tested. It can be concluded that the average four-cell parameters have significant differences.

The fourth hypothesis testing states that, 'there is a difference between the reliability of multiple-choice tests are scored by using penalty scoring technique and which scored by using scoring without correction technique for multiple choices tests 3 options'. The result of calculation is *p*-value = 0.000, at a significance level of 0.05 with H_0 rejection criteria where p-value is lesser than *α*-value. Therefore, *p*-value = 0.000 is lesser than *α* = 0.05, and it was decided that the research hypothesis was tested. This means that there are significant differences in reliability between multiple-choice tests are scored by using penalty scoring technique and by using scoring without correction technique for multiple-choice tests with three options.

The fifth hypothesis testing states that, 'there is a difference between the reliability of multiple-choice tests are scored by using scoring penalty technique and which scored by using scoring without correction technique for multiple-choice tests 4 option'. The result of calculation is *p*-value = 0,000, at a significance level of 0.05 with H_0 rejection criteria where p-value is lesser than *α*-value. Therefore, *p*-value = 0.000 is lesser than *α* = 0.05, and it was decided that the research hypothesis was tested. This means that there are significant differences in reliability between multiple-choice tests are scored by using scoring penalty technique and by using scoring without correction technique for multiple-choice tests with four options.

The sixth hypothesis testing states that, 'there is a difference between the reliability of multiple-choice test that uses three answers options and multiple-choice test that uses four answers options by using scoring penalty technique'. The result of calculation is *p*-value = 0.000, at a significance level of 0.05 with H_0 rejection criteria where p-value is lesser than *α*-value. Therefore, *p*-value = 0.000 is lesser than *α* = 0.05, and it was decided that the research hypothesis was tested. This means that there are significant differences in reliability between multiple-choice test that uses three answer options and multiple-choice test that uses four answer options using scoring penalty technique.

The seventh hypothesis testing states that, 'there is a difference between the reliability of multiple-choice test that uses three answers options and multiple-choice test that uses four answers options by

using scoring without correction techniques'. The result of the calculation is *p*-value = 0.000, at a significance level of 0.05 with H_0 rejection criteria where p-value is lesser than α-value. Therefore, *p*-value = 0.000 is lesser than $\alpha = 0.05$, and it was decided that the research hypothesis was tested. This means that there are significant differences in reliability between multiple-choice tests that use three answer options and multiple-choice test that uses four answer options by using scoring without correction techniques.

Based on the results of the first hypothesis test, it was stated that in the hypothesis it was decided to reject H_0, which means that there was a difference in the reliability scores between multiple-choice tests which were suspended through penalty scoring techniques with and without multiple-choice tests. If viewed from the coefficient aspect of the average cell, it shows that the multiple-choice test that is scored with the penalty scoring technique seems more reliable when compared to the multiple-choice tests which is scored by scoring technique without correction. The empirical findings turned out to be in accordance with the previous concept which found that guessing had an impact on random factors into test scores that reduced reliability and validity (Bereby-Meyer et al., 2002; Burton, 2001; Kubinger et al., 2010; Prihoda et al., 2006). Test developers cannot distinguish between correct answers based on mastery of knowledge versus guess-based answers (Bar-Hillel et al., 2005). Basically, it was decided that the penalty scoring technique resulted in a more stable reliability index compared to the scoring technique without correction.

In testing the second research hypothesis, it is explained that in the hypothesis it was decided to reject H_0, which means that there is a difference in the reliability of the score between multiple-choice tests using 3 answer options with multiple-choice tests using four answer options. If viewed from the coefficient aspect, the average cell shows that multiple-choice tests using three answer options appear to be more stable compared to multiple-choice tests that use four answer options. The empirical findings of this study appear to be in line with the theoretical review presented by Dehnada et al. (2014), stating that the use of three answer choices can save time to include more content and items in a test, so as to improve test validity and reliability. The use of three choices in multiple-choice questions seems easier and less demanding for teachers in developing questions because they will not be forced to use distractors that are unreasonable and disabled in writing items.

The findings of this study are in line with the findings from Berrios et al. (2005) and Rodriguez (2005), which state that in its systematic review it is reported that in most cases, when the number of options decreases, reliability decreases except in the case of a reduction from four to three options where reliability will slightly increase. Tarrant and Ware (2010) reported that the choice of version three multiple-choice test was very likely to contain more functioning distractors. Adding deception to multiple-choice tests is a difficult matter. Impostors who are not good (will not function) will threaten the reliability of the test. Some of these reviews are in line with the findings of this study to conclude that the multiple-choice objective test three options give a more stable portion of reliability compared to the objective four test option on the subject of the elementary school education unit test.

The discussion of the third hypothesis based on the test results that obtained information that the rejection of H_0 is fulfilled and that there is an interaction between scoring techniques and the number of options for multiple-choice objective test reliability. These results indicate that there are differences in each treatment cell, for those controlled by the score technique and those controlled by the number of options showing different variances. The empirical findings in this study appear to be in line with the theoretical review that shows the relationship between scoring techniques and the number of options in multiple-choice test measurement activities. As Tjalla and Fitriani (2017) show, the number of options

greatly effects the amount of penalties given by the test participant's response. This can have an impact on the standard deviation of measurement and make the variance more diverse, thus these conditions lead to interactions between scoring techniques and many options for the reliability of the multiple-choice objective test.

Reviewing the fourth hypothesis empirically from the test results decides to reject H_0, which indicates that there is a significant difference between multiple-choice tests that are scored with a penalty score technique with multiple-choice tests that are scored with a score correction technique for the three answer options. When viewed from the average coefficient of cells, it shows that the average coefficient of multiple-choice tests was scored through the penalty scoring technique for the three answer options and showed a more stable coefficient compared to the average cell coefficient of multiple-choice tests which was scaled with scoring techniques without correction for three options. This is in line with the view of Guiliford (1956) that through guessing tends to reduce the distribution of scores and by decreasing the distribution of the score will have an impact on lowering the reliability of the test in question. If that view is communicated with the empirical findings, then it shows that it is in line due to the existence of penalties, various responses to scores from each item and each participant will appear and this will have an impact on the respondent's score variance and item score variance.

The results of testing the fifth hypothesis empirically decides to reject H_0, which indicates that there is a significant difference between multiple-choice tests which are scored with a penalty score technique with multiple-choice tests which are scored with a score correction technique for four answer options. When viewed from the mean coefficient of cells, it shows that the average coefficient of multiple-choice tests are scaled through the penalty scoring technique for four answer options shows a more stable coefficient compared to the average cell coefficient of multiple-choice tests which are scaled with scoring techniques without correction for four options. This further reinforces the view of Guiliford (1956) that through guessing tends to reduce the distribution of scores and by decreasing the distribution of the score will have an impact on the low reliability of the test in question. In line with this view, it has been understood beforehand that the existence of penalties will vary the response scores of each item and each participant will appear and this will have an impact on the respondent score variance and item score variance. Thus, the achievement of the coefficient stability of a measurement, the application of a score-free correction technique can be used in either the multiple-choice four-choice objective test or the three-choice multiple-choice test.

In testing the sixth hypothesis, the decision was obtained to reject H_0, which means that there is a significant difference in reliability between the groups of multiple-choice tests using three answer options for the penalty score technique with a group of multiple-choice tests using four options for the penalty score technique. When viewed in terms of the average cell, it shows that the coefficient of the average group of multiple-choice test items using three answer options for the penalty score technique is more stable or higher when compared to the multiple-choice test group using four answer options for the penalty score technique.

One characteristic of multiple-choice tests is the presence of distractors. The function of the distractors influences the reliability of multiple-choice tests (Ali et al., 2016). These empirical findings tend to support the statement put forward by Gronlund (1990), which essentially means that it is difficult to choose or determine comparable answer options that function as distractors. This view means that the effort to add an inappropriate option relatively threatens the quality of the items included in the dimensions of reliability. Therefore, through this empirical study, it is recommended that

if the evaluator feels that the addition of options raises a doubt, it is better to only use three answer options.

Furthermore, in the discussion of the seventh hypothesis, the results of previous tests also decided to reject H_0, which means that there was a significant difference in reliability between groups of multiple-choice tests using three answer options for the technique without correction with a group of multiple-choice tests using four options for score technique without correction. In terms of mean cells, information was obtained that the average group cell multiple-choice test items that used three answer options for score techniques without correction were more stable or higher when compared with a group of multiple-choice test items that used four answer options for score techniques without correction. These empirical findings seem to be in line with the views of Dehnada et al. (2014) with the results of research that show that the reliability of the test with three options is higher than the four options.

Learning evaluation is a pedagogical competency that must be mastered by every teacher (Tjabolo & Herwin, 2020). This is very important because in general the quality of education will depend on the quality and competence of teachers (Pujiastuti et al., 2021; Saptono et al., 2021; Wuryandani & Herwin, 2021). Assessment and use of appropriate techniques in the evaluation of learning will determine the significance of the evaluation results carried out for the improvement of the educational process. Therefore, the details of the learning evaluation must be considered carefully, including the use of answer choices and scoring techniques.

Generally, multiple-choice tests are designed with four to five answer choices. The results of the study obtained by Nwadinigwe and Naibi (2013), who show that the number of answer choices significantly effects the level of difficulty, validity and reliability of the test and the use of three answer choices is more recommended than the others. Based on the results of this study it appears that students who work on multiple-choice test groups using three options have different results (significant) with students working on a group of multiple-choice test items that use four options. Likewise, if viewed from coefficient, the reliability also differs from the two treatment groups until it can decide that the multiple-choice test group using three options gives more reliable results compared to the group of multiple-choice tests using four answer options.

4. Conclusion

Based on the hypotheses testing, the results of the research and discussion, the following can be concluded:

1) There is a significant difference in reliability between the multiple-choice test items scored by using penalty scoring techniques and those scored by using scoring without the correction technique.
2) There is a significant difference in reliability between the multiple-choice test of items that uses three answer options and the multiple-choice test of items that use four answer options.
3) There is the interaction between the technique and numerous options on the reliability of multiple-choice objective tests in social subjects of elementary school.
4) There is a reliability difference between the items group of multiple-choice test scored by using the penalty scoring techniques and items group of multiple-choice test scored by using the scoring without correction technique for multiple-choice test with three options.

5) There is a difference in reliability between the items group of multiple-choice tests are scored by using scoring without correction technique for multiple-choice test with four options.

6) There is a significant difference in reliability between the items group of multiple-choice tests that uses three answers options and the item group of multiple-choice tests that use four answer options for the penalty scoring technique.

7) There is a significant difference in reliability between the items group of multiple-choice tests that uses three answers options and the item group of multiple-choice tests that uses four answer options for scoring without correction technique.

Based on some research hypotheses testing, discussion of research results and conclusion from the results of the implementation of this study, it is recommended that some suggestions are the practice of measurement in education, especially related to the use of the multiple-choice objective test instrument, the evaluator should focus on the implementation of scoring techniques to test participants' work, because it affects the reliability of the tests that are given to participants. To achieve maximum reliability, it is recommended to evaluators apply the penalty scoring techniques in scoring participants' work who are subject to multiple-choice tests.

Measurement practices in education, especially those related to the use of multiple-choice objective test instrument, the evaluator should pay attention to the number of answer options that are used for the test device, because it affects the reliability of the tests that are given to participants. To get a maximum reliability index on characteristics of participants in elementary school, it is recommended that evaluators use three answers option in developing multiple-choice test instruments. In developing multiple-choice test instruments, the evaluator should be really careful in presenting answer options, accurate in developing appropriate distractors, be equal and has a function to keep the quality of multiple-choice tests.

Acknowledgments

The researcher would like to thank the principal and teachers who have supported this research activity.

References

Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: Why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning, 16*(1), 1–14. http://doi.org/10.14434/josotl.v16i1.19106

Annie, W. Y., & Chan, A. H. S. (2009). Different methods of multiple-choice test: Implications and design for further research. *Proceedings of the International MultiConference of Engineers and Computer Scientists, 2.* http://doi.org/10.1.1.159.379

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple-choice test: a case study in irrationality. *Mind & Society, 4*(1) 3–12. https://doi.org/10.1007/s11299-005-0001-z

Bereby-Meyer, Y., Meyer, Y., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple-choice tests. *Journal of Behavioral Decision Making, 15,* 313–327. https://doi.org/10.1002/bdm.417

Berrios, G., Rojas, C., Cartaya, N., & Casart, Y. (2005). Effect of the number of options on the quality of EST reading comprehension multiple choice exam. *Paradigma, 26*(1). http://revistas.upel.edu.ve/index.php/paradigma/article/download/4225/2152

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*(1) 41–50. https://doi.org/10.1080/02602930020022273

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Dehnada, A., Nasserb, H., & Hosseinic A. F. (2014). A comparison between three-and four-option multiple choice questions. *Social and Behavioral Sciences, 98,* 398–403. https://doi.org/10.1016/j.sbspro.2014.03.432

Gronlund, N. 1990. *Constructing achievement test*. Prentice-Hall.

Herwin, H., Fathurrohman, F., Wuryandani, W., Dahalan, S. C., Suparlan, S., Firmansyah, F., & Kurniawati, K. (2022). Evaluation of structural and measurement models of student satisfaction in online learning. *International Journal of Evaluation and Research in Education, 11*(1), 152–160. http://doi.org/10.11591/ijere.v11i1.22115

Herwin, H., & Mardapi, D. (2017). An emotion assessment model for elementary school students. *Jurnal Penelitian dan Evaluasi Pendidikan, 21*(1), 80–92. https://doi.org/10.21831/pep.v21i1.14504

Herwin, H., & Nurhayati, R. (2021). Measuring students' curiosity character using confirmatory factor analysis. *European Journal of Educational Research, 10*(2), 773–783. https://doi.org/10.12973/eu-jer.10.2.773

Karandikar, R. L. (2010). On multiple choice tests and negative marking. *Current Science, 99*(8)*,* 1042–1045. https://www.jstor.org/stable/24066110

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: superiority of two solutions and three distractors item format to one solution and five distractors item format. *International Journal of Selection and Assessment, 18*(1), 111–115. https://doi.org/10.1111/j.1468-2389.2010.00493.x

Kurz, T. B. (1999). *A review of scoring algorithms for multiple-choice tests*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. https://eric.ed.gov/?id=ED428076

Lesage, E., Valcke, M., & Sabbe, E. 2013. Scoring methods for multiple choice assessment in higher education –Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation, 39*(1), 188–193. https://doi.org/10.1016/j.stueduc.2013.07.001

Mardapi, D. (2012). *Pengukuran, penilaian dan evaluasi pendidikan [Measurement, assessment and evaluation of education]*. Nuha Medika.

Nájera Catalán, H. E., & Gordon, D. (2020) The importance of reliability and construct validity in multidimensional poverty measurement: An illustration using the multidimensional poverty index for Latin America (MPI-LA). *The Journal of Development Studies, 56*(9), 1763–1783. https://doi.org/10.1080/00220388.2019.1663176

Nwadinigwe, P. I., & Naibi, L. (2013). The number of options in a multiple-choice test item and the psychometric characteristics. *Journal of Education and Practice, 4*(28)*,* 189–196. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.937.1862&rep=rep1&type=pdf

Otaya, L. G., Kartowagiran, B., & Retnawati, H. (2020). The construct validity and reliability of the lesson plan assessment instrument in primary schools. *Jurnal Prima Edukasia, 8*(2), 126–134. https://doi.org/10.21831/jpe.v8i2.33135

Prihoda, T. J., Pinckard, R. N., McMahan, C. A. & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education, 70*(4), 378–386. https://doi.org/10.1002/j.0022-0337.2006.70.4.tb04092.x

Pujiastuti, P., Herwin, H., & Firdaus, F. M. (2021). Thematic learning during the pandemic: CIPP evaluation study. *Cypriot Journal of Educational Science, 16*(6), 2970–3980. https://doi.org/10.18844/cjes.v16i6.6481

Rintayati, P., Lukitasari, H., & Syawaludin, A. (2021). Development of two-tier multiple choice test to assess Indonesian Elementary Students' higher-order thinking skills. *International Journal of Instruction*, *14*(1), 555–566. https://doi.org/10.29333/iji.2021.14133a

Rodriguez, M. C. (2005). Three-options are optimal for multiple-choice items: A metaanalysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Saptono, B., Herwin, H., & Firmansyah, F. (2021). Web-based evaluation for teacher professional program: Design and development studies. *World Journal on Educational Technology: Current Issues, 13*(4), 672–683. https://doi.org/10.18844/wjet.v13i4.6253

Susetyo, B. (2011). *Menyusun tes hasil belajar dengan teori ujian klasik dan responsi butir [Compiling learning outcomes tests using classical exam theory and item response]*. Cakra.

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three- and four options multiple-choice questions in nursing assessments. *Nurse Education Today, 30*(6)*,* 539–543. https://doi.org/10.1016/j.nedt.2009.11.002

Tjabolo, S. A., & Herwin, H. (2020). The influence of teacher certification on the performance of elementary school teachers in Gorontalo Province, Indonesia. *International Journal of Instruction, 13*(4), 347–360. https://doi.org/10.29333/iji.2020.13422a

Tjalla, A., & Fitriani, S. (2017). The effect of multiple choice scoring methods and risk taking attitude toward chemistry learning outcomes. *Journal of e-Learning and Higher Education, 2017*, 11. https://doi.org/10.5171/2017.496122

Wuryandani, W., & Herwin, H. (2021). The effect of the think–pair–share model on learning outcomes of Civics in elementary school students. *Cypriot Journal of Educational Sciences, 16*(2), 627–640. https://doi.org/10.18844/cjes.v16i2.5640

Ziegler, J., & Detje, F. (2013. *Application of empirical methodology to evaluate information fusion approaches* (pp. 1878–1885). Proceedings of the 16th International Conference on Information Fusion. https://ieeexplore.ieee.org/abstract/document/6641234